Building a Common Framework for IIR Evaluation

Mark M. Hall[1] and Elaine Toms[1]

{m.mhall|e.toms}@sheffield.ac.uk
Information School
University of Sheffield
Sheffield S1 4DP, UK

Abstract. Cranfield-style evaluations standardised Information Retrieval (IR) evaluation practices, enabling the creation of programmes such as TREC, CLEF, and INEX, and long-term comparability of IR systems. However, the methodology does not translate well into the Interactive IR (IIR) domain, where the inclusion of the user into the search process and the repeated interaction between user and system creates more variability than the Cranfield-style evaluations can support. As a result, IIR evaluations of various systems have tended to be non-comparable, not because the systems vary, but because the methodologies used are non-comparable. In this paper we describe a standardised IIR evaluation framework, that ensures that IIR evaluations can share a standardised baseline methodology in much the same way that TREC, CLEF, and INEX imposed a process on IR evaluation. The framework provides a common baseline, derived by integrating existing, validated evaluation measures, that enables inter-study comparison, but is also flexible enough to support most kinds of IIR studies. This is achieved through the use of a "pluggable" system, into which any web-based IIR interface can be embedded. The framework has been implemented and the software will be made available to reduce the resource commitment required for IIR studies.

Keywords: evaluation, methodology, interactive information retrieval

1 Introduction

Cranfield-style evaluations standardised Information Retrieval (IR) evaluation practices, and served as the foundation for a host of evaluation programmes including TREC, CLEF, and INEX. These set the pace for evaluating the output from information retrieval systems with a view to improving system performance. Many accomplishments over the past three decades in search systems effectiveness can be linked to these programmes. In parallel, the interactive IR (IIR) research community focused somewhat similar research on the user as a core ingredient in the research. While there is overlap, IIR has additional goals: a)

assess search systems and components of search systems using user-centred evaluation methods typically found in human experimentation and human computer interaction (e.g., [12]), and b) examine user actions and activities – both cognitive and behavioural – to understand how people search for information and which aspect of context (e.g., characteristics of the user, the work environment, situation, etc.) influences the process (e.g. [4, 10]).

While the TREC and CLEF programmes have enjoyed standardised protocols and measures to assess performance and output, and to experimentally compare among systems, the IIR evaluation field has not had that advantage. The TREC and CLEF evaluation programmes specified standard test collections, test topics and sets of expert-assessed relevant items (including training sets) as the minimum ingredients, and a standard way of presenting and comparing the results – the ubiquitous reverse-ranked list of relevant items per topic and additionally aggregated by system and collection. On the other hand, IIR research was and still is researcher driven with non-standard "collections", user-imposed search tasks, and diverse sets of measures to support multiple research objectives. In the midst of all of this is usually a set of participants, a sample of convenience. Thus, it is difficult to compare across studies.

The challenge is two-fold: developing a standard methodological protocol that may service multiple types of IIR evaluations and research, and developing a standard set of meaningful measures that are more than descriptive of the process. In this work, we address the first: we designed, developed, implemented and tested a common research infrastructure and protocol that can be used by the IIR research community to systematically conduct IIR studies. Over time, the accumulated studies will also provide a comprehensive data set that includes both context and process data that may be used by the IR community to test and develop algorithms seated in human cognition and behaviour, and additionally to provide a sufficiently robust, detailed, reliable data set that may be used to test existing measures and develop new ones. This paper describes the rationale and the design of the infrastructure, and its subsequent implementation.

2 Interactive IR Research – Past and Present

Typically IIR research was conducted using a single system in a laboratory setting in which a researcher observed and interacted with a participant [21]. This was a time-consuming, resource exhaustive and labour intensive process [23, 26]. As a result, IIR research used a small number of participants doing a few tasks, which challenged the validity and reliability of the research [11]. In their recent systematic review of 127 IIR studies, Kelly and Sugimoto [13], found extreme variability in IIR studies: from 4 to 283 participants with a mean of 37, and between six and ten task instantiations was typical, although the maximum observed was 56 in a single study.

Similarly what was measured varied significantly; 1533 measures were identified [13]. Clearly the situation has not changed since Yuan and Meadow examined the measures used in 1999 [27], and Tague-Sutcliff in 1992 [21]. The

challenge has been that the same: concepts are not always measured using the same "yardstick" and there is no standard set. For example, in the outcome from the TREC Interactive Track, lab participants used a similar protocol, but the variables tested differed and measurement was not consistent [6]. All of this variability in IIR studies has not allowed for comparison across a series of studies, or the aggregation of data from multiple studies to test hypotheses in large data sets.

The main challenge lies in creating a framework that is sufficiently standardised to enable comparability of evaluation results, while at the same time being flexible enough to be applied to a wide range of experiments and variables in order to ensure its uptake. The matter has been richly discussed by Tague-Sutcliffe [21] who outlined ten key decisions in the research design. Later, first Ingwersen and Jarvelin 2005 [7] and later Kelly's synthesis of IIR [11], synthesized and elaborated on this process. However, the closest we have come to a standard protocol is the set of instruments used by TREC Interactive Track, and a practice of pre- and post-task data capture that has been used more or less consistently.

While the traditional method for IIR experiments has been in-the-lab studies, the web introduced alternatives that reduced cost, enabled 24-7 experimentation, provided for a high degree of external validity, and to an extent automated parts of the experimental setup [17, 18]. One of the first disciplines to adapt research to the Web was psychology. Its Psychological Research on the Web (http://psych.hanover.edu/Research/exponnet.html) continues to provide links to hundreds of web-based surveys and experiments, but this remains simply a list of links. The Web Experiment List (http://www.wexlist.net) is a similar but parallel service that provides links to and descriptions of current and past web experiments.

In 2004, Toms, Freund and Li designed and implemented the WiIRE (Web-based Interactive Information Retrieval) system [24], which devised an experimental workflow process that took the participant from information page through a variety of questionnaires and the search interface. Used in TREC 11 Interactive Track, it was built using typical Microsoft Office desktop technologies, which severely limited its capabilities. Data collection relied on server logs limiting the amount of client-side data that could be collected. The concept was later implanted in a new version using PHP, JavaScript, and MySQL used in INEX2007 [25]. This version still provided the basics in implementation of a web-based experiment, but lacked flexibility in setup and data extraction. More recently, SCAMP (Search ConfigurAtor for experiMenting with PuppyIR) was developed by Renaud and Azzopardi [19] which is used to assess IR systems, but does not include the range of IIR research designs that are typically done. Another development is the experiment system described in [2], but to our knowledge it is not publicly released. Thus in IIR, there is a significant amount of interest and need to develop standard protocols and systematic approaches to data collection. Given the diversity in past studies and inconsistencies in what is collected and how much, there is a significant need to develop an approach.

3 IIR Evaluation Framework

To overcome these limitations the proposed evaluation framework was designed around five core objectives:

1. Provide a systematic way of setting up an experiment or user study that may be intuitively used by students and researchers;
2. Provide a standard set of evaluation measures to improve comparability;
3. Ensure that standard and consistent data formats are used to simplify the comparison and aggregation of studies;
4. Extract a standard procedure for the conduct of IIR studies from past research, so that studies can share a common protocol even if the system, the tasks, and the participant samples are different;
5. Reduce resource commitment in the conduct of such studies.

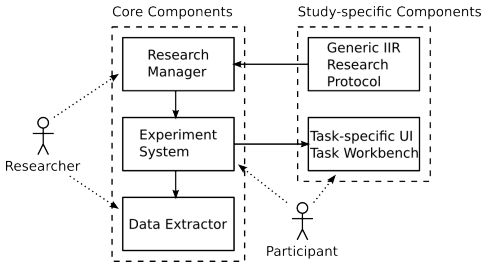

Fig. 1. Design of the proposed evaluation framework, with the three core and the two study-specific components. In a non-IIR study different study-specific components would be used. In the framework, the researcher interacts only with the *Research Manager* and *Data Extractor*, while the participant only ever sees the *Experiment System* and *Task-specific UI*

The difficulty in designing a framework that implements these objectives is balancing the standardisation and simplification efforts with the ability to support the wide range of evaluation experiments conducted within IIR. To achieve this we have developed a flexible framework, inspired by the WiIRE system [24, 25] and work in the POODLE project [2], that provides the core functionality required by all experiments and into which the experiment-specific functionality can easily be plugged-in (fig. 1). The three core components of the framework are:

– The **Research Manager** is the primary point of interaction for the researcher setting up an experiment. It is used to specify the workflow of the experiment, the tasks and interfaces to use, and all other measures to acquire. To simplify and standardise both the experiment process and results, the **Research Manager** is primed with a *generic research protocol*, such as

the *Generic IIR Research Protocol* provided in this paper, that specifies the basic experiment workflow and into which the researcher only has to add the experiment-specific aspects;
- the **Experiment System** takes the experiment defined by the *Research Manager* and generates the UI screens that the participants interact with. It also ensures that the tasks and interfaces are correctly distributed and rotated between the participants, in accordance with the settings specified in the **Research Manager**. Finally it loads the **Task-specific UI** and records the participants' responses and ensures that they conform to the requirements specified by the researcher. To ensure the flexibility of the system, any web-based system can be used as the **Task-specific UI**;
- the **Results Extractor** takes the participant data gathered by the **Experiment system** and provides them in a format that can be used by analysis packages such as SPSS or R. The data includes not only the participants' responses, but also data on tasks / interfaces used by the participants used and the order in which they appeared.

To simplify the setup and further standardise IIR studies, the following two IIR-specific components have been developed. In a non-IIR context, these would be replaced with components developed for that context.

- the **Generic IIR Research Protocol** aims to define a standardised and re-usable workflow and set of evaluation measures for IIR evaluation studies;
- the **Task Workbench** provides an extensible and pluggable set of UI components for IIR interfaces, with the aim of simplifying the set-up of IIR evaluation experiments.

3.1 Research Manager

The *Research Manager* addresses requirements #1 and #5, in that it provides a structured process for setting up experiments and through this reduces the resource commitment required. The *Research Manager* achieves this through the use of *generic research protocol*s that specify a structure for the type of experiment the researcher wishes to conduct. The researcher then adapts this *generic research protocol* to their specific requirements. This provides the desired level of standardisation, while at the same time being flexible enough to support a wide range of experiments. The details will be discussed in the context of IIR evaluation, using the *Generic IIR Research Protocol* in section 3.4, but are equally applicable to any other study that can be conducted via the web.

When setting up an experiment, the researcher first selects the *generic research protocol* that they wish to use, although if there is no applicable *generic research protocol*, then the experiment can also be built from scratch. Assuming the researcher selects the *Generic IIR Research Protocol* to setup an IIR study, they are first asked to provide basic information including title, purpose, key researcher names, and contact information, which are used to generate the initial and final information pages. Next, the researcher selects which of the optional

steps in the *Generic IIR Research Protocol* to include in their study. Naturally this choice can be changed at any time, if testing reveals that optional steps are superfluous or should be included. This specifies the basic structure of the experiment and the next step is to define the core tasks to test or control for, in IIR generally including:

- **Task Type**: categorisation of task based on attributes of a task which may be Fact-finding, Know-item, Topical, Transactional and so on. Unfortunately there is no well-defined taxonomy of task type [22], although multiple types have been created. In this case, Task Type will be defined by the participant, although we hope that current research may provide some parameters around these for greater consistency. Each Task Type, e.g., Topical, is represented by multiple instantiations of that type that specify the exact task that a participant will do using the particular interface and collection. For example, find our who should not get a flu shot. The actual number of task instantiations will vary with the amount of effort that is required of the participant, and this is a decision of the researcher.
- **System**: this may be different IR systems, different interfaces to the same IR system; or a single UI with interface objects.
- **Participant Group**: different groups of participants may be recruited based on selected characteristics. For example, novices may be compared to experts, or youth to seniors, or sometimes by scores on a particular human characteristics such as scores on a cognitive style test.

The researcher first identifies which of these elements will be tested, and whether the design will be between- or within-subjects for *Task Type* and/or *System*, and between-subjects for *Participant Group*. Mixed approaches are also possible to handle scenarios where a pure between- or within-subjects approach is not desired or not feasible. Based on these settings the *Research Manager* creates the final experiment that is then passed onto the *Experiment System*, which then uses the settings to ensure that participants are assigned to *Task type* / *System* / *Participant Group* combinations and that participants are evenly distributed between the combinations.

3.2 Experiment System

The *Experiment System* addresses requirements #3 and #5 by providing a full integrated system that handles the whole workflow of the experiment as it is used by the participants. It takes the experiment designed using the *Research Manager* and guides the participants through the experiment using the three-step workflow shown in figure 2. When a new participant starts the experiment the *Experiment Manager* selects the initial step to show the participant and displays it to the participant. For example, in the *Generic IIR Research Protocol* this is the information and consent form. The participant reads the instructions on the page and answers any questions. They then submit their answers back to the system, which validates the answers against the answer schema defined

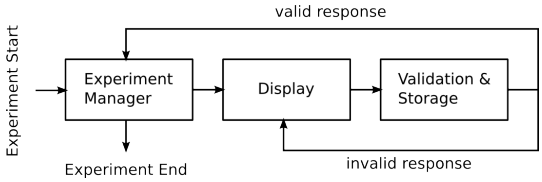

Fig. 2. The main loop implementing the *Experiment System*. Before showing the first step and then after each step the *Experiment Manager* determines the next step, based on the experiment workflow defined in the *Research Manager*, the steps seen so far, and the participant's answers.

in the *Research Manager*. If the results do not match the schema, for example if a required question was not answered, or if the answer is invalid, then the applicable error messages are generated and the page show to the participant again, with their existing answers pre-filled. If the results are acceptable, then the answers are stored and the *Experiment Manager* uses the workflow defined in the *Research Manager* to determine which step to show next. This decision can take into account which steps the participant has completed, which *Task type / System* combination they were assigned to, and also what answers the participant has provided so far.

To ensure that the *Experiment System* can be used in a wide range of experiments, it does not itself include the task interface. At the *Task* steps in the experiment workflow, it simply loads the applicable task UI, as defined in the *Research Manager*, into the interface. A number of different techniques for the embedding are available, including an inline-frame-based, a simple re-direction-based, and a API-callback-based approach. This ensures that the framework can be deployed with most types of web-based UIs and can thus be widely used.

3.3 Data Extractor

The *Data Extractor* addresses requirements #3 in that it outputs the results from the experiment in a standardised format for further processing in analysis packages such as SPSS or R. In addition to the data acquired from the participants, the output also includes data on the *Task type / System* combinations the participants were shown. Simple post-processing steps, such as filtering columns or participant answers, can be applied to the data to reduce the amount of pre-processing required before loading the data into the analysis package.

3.4 Generic IIR Research Protocol

The *Generic IIR Research Protocol* supports requirements #2, #3, and #5 for the IIR evaluation context. By providing a standardised set of steps, ordering of those steps, and measures within those steps, it ensures that results from different studies become comparable. Because the standardised measures are pre-defined, it also reduces the resource commitment required to set up the experiment. To

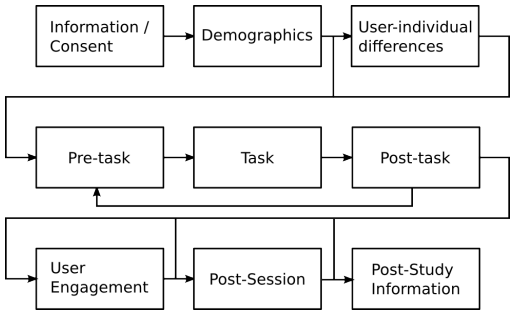

Fig. 3. The main work-flow through the *Generic IIR research Protocol*, showing the optional *user-individual differences* and *post-session* steps and also that the *pre-task* – *task* – *post-task* structure can be repeated multiple times within an experiment if the aim is to evaluate multiple tasks.

be able to support the varied IIR evaluation landscape, it makes no constraint on the IIR UI that is under test, and it also allows the researcher to augment the process with the specific research questions they are interested in (in the *post-task* or *post-session* steps). The protocol has adapted and augmented the protocols used by early TREC Interactive Tracks and INEX Interactive Tracks, all of which are based on many earlier IIR studies. Some aspects have been extracted from more recent work. The main work-flow through the protocol is shown in figure 3 and consists of nine steps:

1. **Study information and Consent**: this is the typical introduction to a study together with a consent form that enables informed consent to be made (which is now expected and required for human-based experiments) and advises participants of their rights in participating. Most of the actual textual content is provided by the researcher when setting up the experiment in the *research manager*. However, because the basic protocol has received Research Ethics approval by Sheffield University, some of the content cannot be modified.
2. **Demographics Questionnaire**: a standard set of questions asked of all participants is used to create a profile of the set of participants in a study. A minimum set of standard variables is required (gender, age, education, cultural background, and employment) to ensure comparability across studies, and in some instances may help explain results (e.g., inexperienced, mostly of one gender, mostly undergraduates and so on). But additional experiment-specific variables can be added to the default set in the *research manager*;
3. **User-individual Differences**: depending on the study objectives, there is a large variety of user characteristics that one might observe, control or test, such as Cognitive Style [20], Need for Cognition [3], Curiosity [9], and Openness to Experience [14]. The basic research protocol does not include any of these as a default; we need more research to emphatically determine that any of these are core predictors of search actions and outcomes. The

Generic IIR Research Protocol defines a standard template to insert these into the experiment, but they will in the short term be study-specific. This customisation is available through the *Research Manager* which may be used to add scales or questions that are not currently specified by the protocol;

4. **Pre-Task Questions**: prior to assigning a participant to a task, the knowledge, experience and interest in the task topic is collected. For this, a set of standard questions derived from TREC and INEX interactive track protocols as well as other IIR studies was used [1]. These will be required, enabling the future comparison across studies. Unlike the implementation in TREC and INEX, the questions have been converted to standard Likert scales requesting agreement with statements;

5. **Task**: at this point in the procedure, the participants are shown the task UI. The UI may be created using our *Task Workbench* or the UI to any web-based system may be inserted. The system used is not discussed further in this paper, as search interfaces is a different topic. The system also handles the insertion of tutorial, and practice in the case of novel interfaces for which a participant may require training and some exposure;

6. **Post-Task Questions**: as with the *pre-task questions* a set of post-task questions also derived from past TREC and INEX interactive tracks, and reproduced in other studies, are integrated into the research protocol as a required step. These questions address the user-perception of completing the assigned task.

7. **User Engagement**: after completing all tasks, a set of post-session questions assesses the participants' engagement with the whole study. By default the *generic research protocol* provides the User Engagement Scale [15]. This scale measures six components of user experience, namely Focused Attention, Perceived Usability, Aesthetics, Endurability, Novelty, and Felt Involvement. At present, there is no competitor for this measure. While we recommend that it be included so that the scale can be further generalised and potentially improved, it is not a required feature.

8. **Post Study**: an additional but not required feature is the option of assessing the interface to the system used and/or the content. However researchers may substitute specific questions aimed at evaluating the whole session. For example in studies testing a novel IIR interface or component, questions evaluating the participants' interactions with the novel interface or component would be asked at this point.

9. **Post Study Information**: minimally this will contain acknowledgement and contact information. Optionally, the participants will also be able to sign up for future studies, with the goal of building up a pool of potential participants for future IIR evaluations. In this case, the system will collect contact information and a brief profile so that targeted recruitment may be conducted.

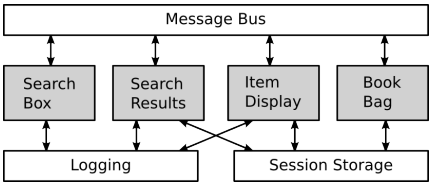

Fig. 4. The pluggable task work-bench provides three shared modules (*Message Bus*, *Logging*, and *Session Storage* into which the actual evaluation UI components (sample shown in grey) are plugged.

3.5 Task Workbench

To further reduce the resource commitments (requirement #5) required to set up an IIR evaluation experiment, an extensible, pluggable task work-bench is provided (fig. 4). The task work-bench provides three standard modules (*Message Bus*, *Logging*, *Session Storage*) into which the experiment / task-specific components are plugged. Each component defines a set of messages it can send and listen for. The researcher then specifies which components should listen to which messages from which other components and the *message bus* ensures that the messages are correctly delivered. This means that new components can easily be integrated with existing components, simply by linking them via their messages.

```
{"participant": 322, "timestamp": "2013-02-13T14:34:23",
 "action": "query", "parameters": {"q": "Railwy"},
 "components": {
   "search_box": {"spelling": "Railway", "q": "Railwy"},
   "search_results": {"numFound": 4, "docs": [{...}, {...}]}}
```

Fig. 5. Example entry for the log-file generated by the *Task Workbench*. The entry shows that participant 322 sent a query "Railwy", together with a list of those components that reacted to the query and what data they showed the participant.

The *Task Workbench* provides standard *logging* and *session storage* modules to simplify the creation of new components. A set of standard components (search box, search results, item display, task display, book-bag for collecting items) that can be re-used or extended. It also generates a very rich log file (fig. 5). In addition to the standard fields it also includes detailed information on which UI components were updated based on the request, and all the data that the updated UI components displayed to the participant. This makes it possible to fully re-play the participant's interaction with the system.

4 Conclusion

In this paper we present a novel, standardised design and system for Interactive Information Retrieval (IIR) experiments, building on past implementations [21, 23, 18, 2]. The framework defines a standardised set of questions that enables the comparability of IIR evaluation results, while still being flexible enough to allow for the investigation of experiment-specific research questions. To reduce the resource requirements of setting up IIR evaluations the framework is supported through a number of extensible software components, that can easily be integrated with existing IIR systems. The goal of the framework is to achieve a level of standardisation in IIR that extends the comparability that Cranfield-style evaluation brought to IR in general to the IIR evaluation domain.

The system has successfully been deployed for the data-collection in the 2013 CLEF CHiC Interactive task [16] and also in the 2013 TREC Session Track [8]. It has also been used in non-IIR studies [5].

5 Acknowledgements

The research leading to these results was supported by the Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191. The authors gratefully acknowledge the advice and contributions of V. Petras, B. Larsen, and P. Hansen to the design.

References

1. Trec 2002 interactive track guidelines. Technical report, 2002.
2. R. Bierig, J. Gwizdka, and M. Cole. A user-centered experiment and logging framework for interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and interpreting user interactions in information search and retrieval*, pages 8–11, 2009.
3. J. T. Cacioppo, R. E. Petty, and C. F. Kao. The efficient assessment of need for cognition. *Journal of personality assessment*, 48(3):306–307, 1984.
4. J. Gwizdka. Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology*, 61(11):2167–2187, 2010.
5. M. Hall, P. Clough, and M. Stevenson. Evaluating the use of clustering for automatically organising digital library collections. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 323–334. Springer Berlin / Heidelberg, 2012. 10.1007/978-3-642-33290-6_35.
6. W. Hersh. Trec 2002 interactive track report. In *Proc. TREC*, 2002.
7. P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer, 2005.
8. E. Kanoulas, M. Hall, P. Clough, and B. Carterette. Overview of the trec 2013 session track. In *The Twentieth Text REtrieval Conference (TREC 2013) Proceedings*, 2013.

9. T. B. Kashdan, M. W. Gallagher, P. J. Silvia, B. P. Winterstein, W. E. Breen, D. Terhar, and M. F. Steger. The curiosity and exploration inventory-ii: Development, factor structure, and psychometrics. *Journal of research in personality*, 43(6):987–998, 2009.
10. D. Kelly. Measuring online information seeking context, part 1: background and method. *Journal of the American Society for Information Science and Technology*, (14):1862–1874, 2006.
11. D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1):1–224, 2009.
12. D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 371–378. ACM, 2009.
13. D. Kelly and C. Sugimoto. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *JASIST*, 64(4):745–770, 2013.
14. K. Lee and M. Ashton. The hexaco personality inventory: A new measure of the major dimensions of personality. *Multivariate Behavioral Research*, 39:329–358, 2004.
15. H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2009.
16. V. Petras, M. Hall, J. Savoy, T. Bogers, P. Malak, E. Toms, and A. Pawlowski. Cultural heritage in clef (chic) 2013.
17. U.-D. Reips. Standards for internet-based experimenting. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, 49(4):243–256, 2002.
18. U.-D. Reips and R. Lengler. Theweb experiment list: A web service for the recruitment of participants and archiving of internet-based experiments. *Behavior Research Methods*, 37(2):287–292, 2005.
19. G. Renaud and L. Azzopardi. Scamp: a tool for conducting interactive information retrieval experiments. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 286–289. ACM, 2012.
20. R. J. Riding and S. Rayner. *Cognitive styles and learning strategies: Understanding style differences in learning and behaviour*. D. Fulton Publishers, 1998.
21. J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992.
22. E. Toms. *Task-based information searching and retrieval*, pages 43–59. Facet Publishing, 2011.
23. E. G. Toms, L. Freund, and C. Li. Wiire: the web interactive information retrieval experimentation system prototype. *Information Processing & Management*, 40(4):655–675, 2004.
24. E. G. Toms, L. Freund, and C. Li. Wiire: the web interactive information retrieval experimentation system prototype. *Information Processing & Management*, 40(4):655–675, 2004.
25. E. G. Toms, H. OBrien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. Macnutt. Task effects on interactive search: The query factor. In *Focused access to XML documents*, pages 359–372. Springer, 2008.
26. E. G. Toms, R. Villa, and L. McCay-Peet. How is a search system used in work task completion? *Journal of Information Science*, 39(1):15–25, 2013.
27. W. Yuan and C. T. Meadow. A study of the use of variables in information retrieval user studies. *Journal of the American Society for Information Science*, 50(2):140–150, 1999.